Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Sequencing Platform Modeling and Analysis

Nonparametric Tests for Differential Histone Enrichment with ChIP-Seq Data

Qian Wu, Kyoung-Jae Won and Hongzhe Li

Department of Biostatistics and Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA.

ABSTRACT: Chromatin immunoprecipitation sequencing (ChIP-seq) is a powerful method for analyzing protein interactions with DNA. It can be applied to identify the binding sites of transcription factors (TFs) and genomic landscape of histone modification marks (HMs). Previous research has largely focused on developing peak-calling procedures to detect the binding sites for TFs. However, these procedures may fail when applied to ChIP-seq data of HMs, which have diffuse signals and multiple local peaks. In addition, it is important to identify genes with differential histone enrichment regions between two experimental conditions, such as different cellular states or different time points. Parametric methods based on Poisson/negative binomial distribution have been proposed to address this differential enrichment problem and most of these methods require biological replications. However, many ChIP-seq data usually have a few or even no replicates. We propose a nonparametric method to identify the genes with differential histone enrichment regions even without replicates. Our method is based on nonparametric hypothesis testing and kernel smoothing in order to capture the spatial differences in histone-enriched profiles. We demonstrate the method using ChIP-seq data on a comparative epigenomic profiling of adipogenesis of murine adipose stromal cells and the Encyclopedia of DNA Elements (ENCODE) ChIP-seq data. Our method identifies many genes with differential H3K27ac histone enrichment profiles at gene promoter regions between proliferating preadipocytes and mature adipocytes in murine 3T3-L1 cells. The test statistics also correlate with the gene expression changes well and are predictive to gene expression changes, indicating that the identified differentially enriched regions are indeed biologically meaningful.

KEYWORDS: kernel smoothing, normalization, nonparametric testing, spatial histone profiles

SUPPLEMENT: Sequencing Platform Modeling and Analysis

CITATION: Wu et al. Nonparametric Tests for Differential Histone Enrichment with ChIP-Seq Data. *Cancer Informatics* 2015:14(S1) 11–22 doi: 10.4137/CIN.S13972.

RECEIVED: April 10, 2014. **RESUBMITTED:** August 31, 2014. **ACCEPTED FOR PUBLICATION:** September 3, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Methodology

FUNDING: This research was supported by NIH grants CA127334 and GM097505. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: hongzhe@upenn.edu

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties.

Introduction

Chromatin immunoprecipitation sequencing (ChIP-seq) technology is a powerful tool for analyzing protein interactions with DNA.[1] ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins. It can be used to map global binding sites of transcription factors (TFs) and genomic landscape of histone modification marks (HMs). This high-throughput technology creates millions of short parallel sequencing reads and provides more accurate mapping information for the binding regions in the whole genome with lower cost[2–5] than array-based methods. Both TF binding and histone modification play important roles in gene regulation, where TFs bind to DNA at a promoter region to promote or block gene transcription. The signal of TFs usually shows one sharp peak at their binding sites. Multiple HMs have been reported to be associated with transcription initialization, open chromatin, and repression of transcription.[3,6]

Most previous works in analysis of ChIP-seq data have focused on developing peak-calling procedures to find the binding sites for TFs.[7–11] Identifying the enriched regions of HMs is difficult since their signals are more spread out.[12] The signals of HMs are diffuse and usually have multiple local

peaks, which are hard to identify by directly applying the existing peak-calling algorithms.

Another important question is to identify the genomic regions that show differential enrichment of histone modification between two experimental conditions, such as different cellular states or different time points.[3,13] Indeed, different types of differential histone enrichment have been observed, including shift of nucleosome positions, peak height differences and presence/absence of HMs.[14,15] Chen et al.[14] further demonstrated that the spatial distributions of histone marks are predictive for promoter locations and promoter usage. Angel et al.[16] showed that during cold, the H3K27me3 levels progressively increase at a tightly localized nucleation region in *Arabidopsis*, indicating the importance of studying the peak height, not just presence/absence of the peaks.

One common approach to identifying differentially enriched regions by histones is to apply a peak-calling algorithm to identify the enriched regions for each of the two conditions. The regions with peaks in one condition but without peaks in the other condition are then selected. However, selection of enriched regions often depends on the thresholds used in the peak-calling algorithm. Small differences in the calculated *P* values or the False Discovery Rate (FDR) threshold used by the peak-finding procedures can lead to very different sets of peaks. Furthermore, this simple procedure has limitations in detecting the differential enrichment in terms of different peak heights or different peak locations.

Several parametric methods based on Poisson/negative binomial distribution have been proposed to address this differential enrichment problem in ChIP-seq data such as DiffBind and DBChIP.[13,17] Most of these methods require biological replications to estimate the parameters, especially the dispersion parameter in the negative binomial model.[8] However, many ChIP-seq data usually have a few or even no replicates. Taslim et al.[18] proposed a nonlinear method that uses locally weighted regression (Lowess) for ChIP-seq data normalization. Shao et al.[19] developed a method to quantitatively compare ChIP-seq data sets. To circumvent the issue of differences in signal-to-noise ratios between samples, they focused on ChIP-enriched regions and introduced the idea that ChIP-seq common peaks could serve as a reference to build the rescaling model for normalization. The inputs of all the methods mentioned rely on first identifying the enriched regions and then obtaining the total tag or read counts in these regions. Such approaches have two limitations. First, one has to identify the regions using peak-finding algorithms. Second, by summarizing the number of tags into one single number of the region, one can potentially lose important spatial profile differences such as shifts of the signal region or shapes of signals.

In this paper, we propose a nonparametric method to identify the genes with differentially enriched regions based on the ChIP-seq data of histones. Instead of first identifying the enriched regions or peaks as most of the existing methods do, we consider the regions close to genes that may contain important regulatory elements such as the promoter regions, the gene body, and downstream regions of the genes. For each of these regions, we summarize the data as counts of sequencing reads in each of the bins of a given length (eg, 25 bps). The counts in these candidate regions provide important information about different HM enrichment levels between two cellular states. After transforming the count data to approximately normal, we apply kernel smoothing to the differences of the data and develop a nonparametric hypothesis testing procedure based on the kernel smoothing. Applying smoothing to the data helps to eliminate the small local differences that are unlikely to be biologically relevant.

We demonstrate the method using ChIP-seq data on a comparative epigenomic profiling of adipogenesis of murine 3T3-L1 cells. Our method detects genes with differential H3K27ac levels at gene promoter regions between proliferating preadipocytes and mature adipocytes, which agree with what were observed by Mikkelsen et al.[3] The test statistics correlate with the gene expression changes well, indicating that the identified differences are indeed biologically meaningful. Our results also indicate that the combination of different histone modification profiles can predict the fold changes of gene expressions very well.

Motivating Comparative ChIP-Seq Study, Data Transformation, and Statistical Model

We consider the ChIP-seq experiments reported by Mikkelsen et al.[3] on murine 3T3-L1 cells undergoing adipogenesis. Specifically, they generated genome-wide chromatin state maps using ChIP-seq profiling, where they mapped six HMs and two TFs at four time points, including proliferating (day –2) and confluent (day 0) preadipocytes, immature adpipocytes (day 2), and mature adipocytes (day 7). We focus our analysis on H3K27ac mark, which is expected to be enriched at active promoters or enhancers. In order to identify the genes that show differential H3K27ac modification levels between the preadipocytes (day –2) and mature adipocytes (day 7), we consider the upstream 5000 bp region and downstream 2000 bp region around the transcription start site for each gene and divide the regions into 280 bins of 25 bps. We map the raw data using Bowtie,[20] extend reads to the fragment size and then obtain the genome-wide coverage data with a fixed bin size of 25 bp. Since the two ChIP-seq samples are usually sequenced at different depths (total number of reads), we rescale the counts according to the sequencing depth ratio. Suppose that there are m genes and for each gene i, there are n observed, we have read counts X_{ikj} in bin k under condition j, for $i = 1, \ldots, m$; $k = 1, \ldots, n$; and $j = 1, 2$. Our goal is to identify the genes with differential H3K27ac levels at their promoter regions between mature adipocytes and preadipocytes.

For each gene i and each condition j, we assume the data X_{ikj}, $k = 1, \ldots, n$ are approximately Poisson with means μ_{ikj}. We first apply variance-stabilizing transformation (VST)

procedure to transform the data to $X^*_{ikj} = 2\sqrt{X_{ikj} + 0.25}$, as recommended by Brown et al.[21,22] We then treat X^*_{ikj}'s as approximate normal random variables with mean $2\sqrt{\lambda_{ikj}}$ and variance of 1. For the ith gene, in order to test for differential enrichment between the two conditions, we calculate the difference between the two conditions as $Y_{ik} = X^*_{ik1} - X^*_{ik2}$. If there is no differential enrichment, $Y_i^T = (Y_{ik,\ldots,} Y_{in})$ should have a mean value of zero.

We further denote $Y_i(t_k) = Y_{ik}$, for $t_k = k/n \in (0, 1]$, $k = 1, \ldots, n$. We assume the following "signal+white noise" model for the count differences after the VST,

$$Y_i(t_k) = f_i(t_k) + \sigma_i W_i(t_k), \tag{1}$$

where $f_i(t)$ is a smooth function that characterizes the difference of the ChIP-seq enrichment profiles, $W_i(t_k)$ is Gaussian noise with mean 0 and variance 1, and σ_i^2 is the noise variance. For the ith gene, the null hypothesis that there is no differential enrichment between the two conditions is equivalent to testing

$$H_0 : f_i(t) = 0. \tag{2}$$

Kernel-Smoothing-Based Nonparametric Tests

For a given gene i, we propose a kernel-smoothing based nonparametric test[23] to test the null hypothesis (2). For notational simplicity, we omit the subscript i in the following. Let K be a proper kernel, which is a symmetric, continuous density function with an expectation of zero. We use a normal kernel function, which satisfies all these regularity conditions and fits the real data well. For a fixed bandwidth value $\lambda \in [0, 1]$, we consider the kernel estimator $\tilde{Y}_\lambda(t)$ with $t \in [0, 1]$, $s \in [0, 1]$ and its standard decomposition as

$$\begin{aligned}\tilde{Y}\hat{\lambda}(t) &= \frac{1}{\lambda}\int_0^1 K\left(\frac{t-s}{\lambda}\right)Y(s)ds \\ &= \frac{1}{\lambda}\int_0^1 K\left(\frac{t-s}{\lambda}\right)f(s)ds + \frac{\sigma}{\lambda}\int_0^1 K\left(\frac{t-s}{\lambda}\right)W(s)ds \\ &= f\hat{\lambda}(t) + \sigma\xi_\lambda(t),\end{aligned} \tag{3}$$

where

$$f_\lambda(t) = \frac{1}{\lambda}\int_0^1 K\left(\frac{t-s}{\lambda}\right)f(s)ds \text{ and } \xi_\lambda(t) = \frac{1}{\lambda}\int_0^1 K\left(\frac{t-s}{\lambda}\right)W(s)ds.$$

Based on the study by Lepski and Spokoiny,[23] we use the integral of the squared kernel estimator T_λ, which is defined as

$$T_\lambda = \frac{||\tilde{Y}_\lambda||^2}{\hat{\sigma}^2} = \frac{\int_0^1 \tilde{Y}_\lambda^2(t)dt}{\hat{\sigma}^2} \tag{4}$$

to test the null hypothesis H_0: $||f(t)|| = 0$, where $\hat{\sigma}^2$ is some estimate of the error variance, which we discuss in the Estimate σ for Each Gene section. Under the null H_0, one has

$$\hat{Y}_{0\lambda}(t) = \sigma\xi_\lambda(t) \tag{5}$$

and the test statistic becomes $T_{0\lambda} = \int_0^1 \xi_\lambda^2(t)dt$. Since $W(t_i)$ follows $N(0, 1)$, we have

$$\xi_\lambda(t) = \frac{1}{\lambda}\int_0^1 K\left(\frac{t-s}{\lambda}\right)W(s)ds.$$

For the Gaussian kernel, the expectation of $T_{0\lambda}$ is given by

$$E(T_{0\lambda}) = \frac{1}{n\lambda}||K||^2 = \frac{1}{n\lambda}\frac{1}{2\sqrt{\pi}},$$

and its variance is

$$Var(T_{0\lambda}) = \frac{1}{n^2\lambda}\frac{1}{\sqrt{2\pi}}.$$

We define the test statistic as

$$Z_{0\lambda} = \frac{T_\lambda - E(T_{0\lambda})}{\sqrt{Var(T_{0\lambda})}}, \tag{6}$$

which follows $N(0, 1)$ as $n \to \infty$ under the null hypothesis.

Alternative derivation of the test statistic. In this section, we present an alternative derivation of the test statistic that has better finite sample performance than the statistic (6) when n is not too large (see the Application to a Comparative ChIP-Seq Study During Mouse Adipogenesis section for an illustration). Note that the kernel smoother $\tilde{Y}_\lambda(t)$ can be written as a linear combination of $Y^T = (Y_1, \ldots, Y_n)$,

$$\tilde{Y}_\lambda(t) = S_\lambda Y, \tag{7}$$

where S_λ is considered as the hat matrix,

$$S_\lambda = \frac{1}{n\lambda}\begin{pmatrix} K\left(\frac{t_1-s_1}{\lambda}\right) & \cdots & K\left(\frac{t_1-s_n}{\lambda}\right) \\ \vdots & \ddots & \vdots \\ K\left(\frac{t_n-s_1}{\lambda}\right) & \cdots & K\left(\frac{t_n-s_n}{\lambda}\right) \end{pmatrix}.$$

The trace of S_λ is the degrees of freedom (df) of the kernel smoother.[24]

Based on equations (3), (4), and (7), the statistic T_λ can be approximated by

$$T_\lambda = \frac{1}{n\sigma^2}\sum_{k=1}^n \tilde{Y}_{k\lambda}^2 = \frac{1}{n\sigma^2}Y^T S_\lambda^T S_\lambda Y, \tag{8}$$

where the $n \times n$ matrix S_λ^T is the transpose of S_λ. Let $M = S_\lambda^T S_\lambda$ with the following eigen decomposition, $V^T M V = D$, where

$D = diag(d_1, \ldots, d_n)$, $d_1 \geq \ldots \geq d_n$, are the eigenvalues and V is the orthogonal matrix of the eigenvectors. Under the null hypothesis, based on equation (5), Y/σ follows a multivariate normal distribution $N_n(0, I_n)$. Let $U^T = (U_1, \ldots, U_n) = V^T Y/\sigma$, we can rewrite T_λ as

$$T_\lambda = \frac{1}{n} U^T D U = \frac{1}{n} \sum_{k=1}^{n} d_k U_k^2.$$

Since V is an orthogonal matrix, the vector U follows $N_n(0, VV^T) = N_n(0, I_n)$ under the null hypothesis, and therefore U_k^2 are *i.i.d* random variables following χ_1^2 and T_λ follows a mixture of n χ^2 distributions with weights d_k/n. Furthermore, based on the study by Bentler and Xie,[25] under the null hypothesis, T_λ can be approximated by a weighted χ^2 distribution, $\delta\chi_d^2$, where

$$d = \frac{\left(\sum_{k=1}^{n} d_k\right)^2}{\sum_{k=1}^{n} d_k^2}, \quad \delta = \frac{\left(\sum_{k=1}^{n} d_k\right)}{d}.$$

Alternatively, using the Wilson–Hilferty transformation,[26] we have

$$Z_{0\lambda,WH} = \frac{\sqrt[3]{\frac{T_\lambda}{\delta d}} - \left(1 - \frac{2}{9d}\right)}{\sqrt{\frac{2}{9d}}}, \tag{9}$$

which follows a $N(0, 1)$ under the null hypothesis. We use this statistic in our analysis.

Estimate σ for each gene. In order to calculate the test statistic specified as equation (4) or (8), we need the variance estimate $\hat{\sigma}_i^2$ for each gene *i*. After the VST of the read counts, for each gene *i*, we assume that the observations Y_{ik} have the same variance σ_1^2. We consider the Nadaraya–Watson nonparametric regression with kernel smoothers as (3),

$$\tilde{Y}_\lambda(t) = S_\lambda Y,$$

where $df = tr(S_\lambda)$ is the degrees of freedom of the kernel smoother.[24] We estimate the variance σ_i^2 by calculating the residual sum of squares

$$\hat{\sigma}^2 = \frac{\left[\tilde{Y}_\lambda(t) - Y(t)\right]^T \left[\tilde{Y}_\lambda(t) - Y(t)\right]}{n - df} = \frac{\Sigma_{k=1}^{n} \left[Y_k - \tilde{Y}_\lambda(t_k)\right]^2}{n - df}. \tag{10}$$

Since we consider the ChIP-seq data with very few or no replications, the estimates $\hat{\sigma}_i^2$ can be too small for very small counts. To improve precision, we use an approach similar to that used by Efron et al.[27] and Tusher et al[28]: we add a constant a_0 = 90th percentile of the standard deviations to make the standard deviation of each gene bigger to avoid false identification of genes with differential enrichment. The final modified estimator of the variance is $\hat{\sigma}_i^2(\hat{\sigma}_i + a_0)^2$.

Finally, we choose the bandwidth λ in the kernel smoothing to be relatively large to avoid fitting the very small local changes. In our analysis of the real data sets with $n = 280$ observations, we choose $\lambda = 20/280$. The details of bandwidth selection are discussed in the Effects of Bandwidth Selection on Identifying the Differential Enrichment Genes section.

Application to a Comparative ChIP-Seq Study During Mouse Adipogenesis

We present results of our analysis of the comparative ChIP-seq data described in the Motivating Comparative ChIP-seq Study, Data Transformation and Statistical Model section. Our initial analysis focuses on H3K27ac at gene promoter regions, because it is known that H3K27ac is positively associated with gene expression.[3] We divide the genomic region around the transcription starting site (–5000 to 2000 bp) into $n = 280$ bins, where the length of each bin is 25 bps. The data set includes $m = 29{,}716$ genes. Our goal is to identify the genes with differential enrichment of H3K27ac at the promoter regions between proliferating preadipocytes (day –2) and mature adipocytes (day 7).

Comparison of the $Z_{0\lambda,WH}$ statistics and fold-change statistics. For each gene, after the normal transformation as in the Motivating Comparative ChIP-seq Study, Data Transformation and Statistical Model section, we fit a kernel-smoothing function to the difference data using a bandwidth of $\lambda = 20/280$, which over-smooths the very small signals that are likely due to noises. We calculate the test statistic for each of the 29,716 genes. To compare different test statistics $Z_{0\lambda}$ and $Z_{0\lambda,WH}$, we plot the histograms of these two test statistics in Figure 1 for 9,874 genes with the maximum number of read counts in both days fewer than 5. Because of the very small read counts in these genes, these genes are most likely not differentially enriched and therefore the test statistics should follow the standard normal distribution. Clearly, $Z_{0\lambda,WH}$ follows $N(0, 1)$ closer than $Z_{0\lambda}$. We therefore use this statistic in all our analyses.

Using the test statistic $Z_{0\lambda,WH}$, we observed that about one-third of the genes that show differential enrichment between preadipocytes and mature adipocytes using a Bonferroni-adjusted *P*-value of 0.05. This is expected since the cells are very different between these two days. Large-scale differential enrichment was also observed by Mikkelsen et al.[3] We observe different patterns of differential enrichment. Figure 2 shows the observed data for 12 genes with the largest test statistics. Clearly, some genes are enriched for H3K27ac in only one condition. For genes that are enriched at both time points, Figure 2 shows that these genes have different H3K27ac enrichment levels or peak heights.

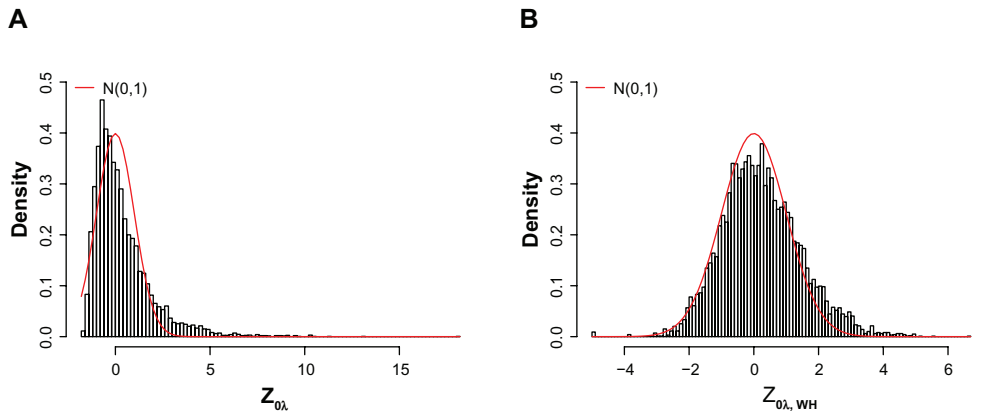

Figure 1. Histograms of two test statistics for the mouse adipogenesis ChIP-seq data, (**A**) $Z_{0\lambda}$ and (**B**) $Z_{0\lambda,HW}$, for 9,874 genes with the maximum number of read counts in both day –2 and day 7 fewer than 5. The red curve in each plot represents the standard normal density.

As a comparison, for each of the genes, we also calculate the simple fold-change statistics and the statistics used in DBChIP.[13] In general, we observe that large $Z_{o\lambda,HW}$ statistics correspond to large fold changes or large DBChIP statistics. We observe a small set of genes that have very small $Z_{0\lambda,WH}$ statistics, but with very large fold changes or DBChIP statistics. These genes tend to have very small read counts. We also observe that some genes have very small fold changes, but with large $Z_{0\lambda,HW}$ statistics. Figure 3 shows the plots of 12 such genes. Many such genes show a clear shift of peaks between two different cell states, which cannot be captured simply using total read counts as in fold changes and the DBChIP statistics. This indicates the importance of modeling the spatial ChIP enrichment profiles.

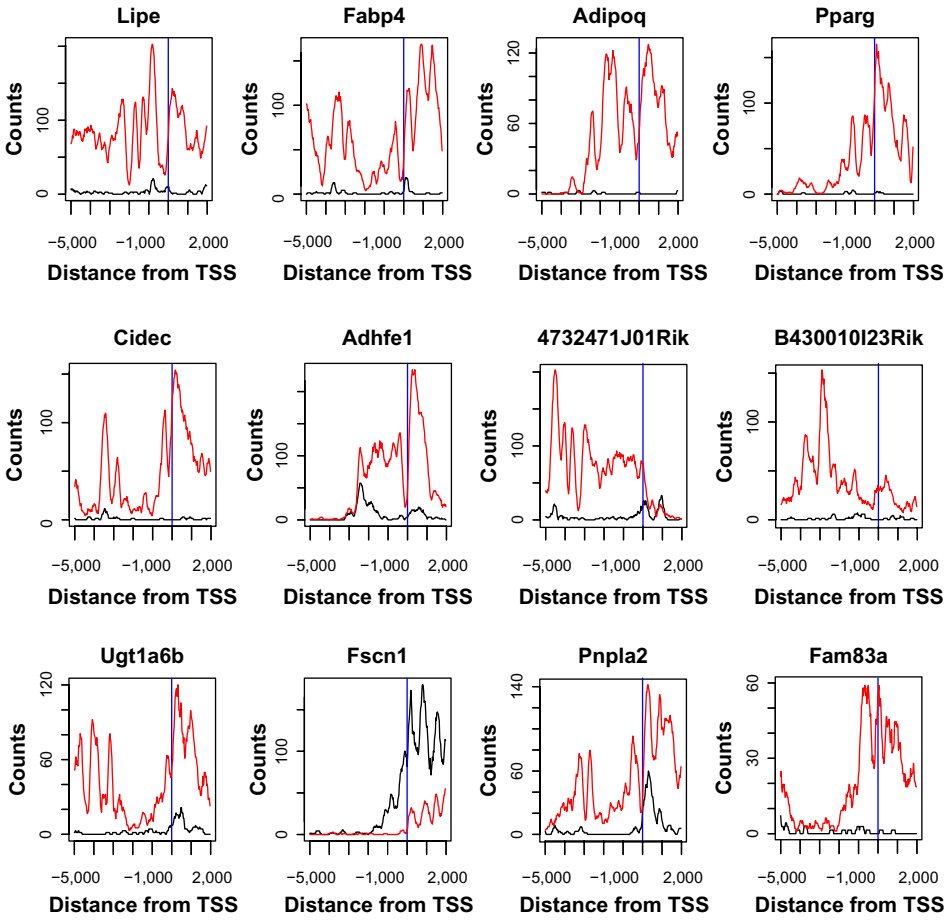

Figure 2. Observed mouse adipogenesis ChIP-seq bin-counts for top 12 genes ranked by the test statistics $Z_{0\lambda,WH}$ over the promoter region for day –2 (red) and day 7 (black). Vertical line represents the transcription starting site.

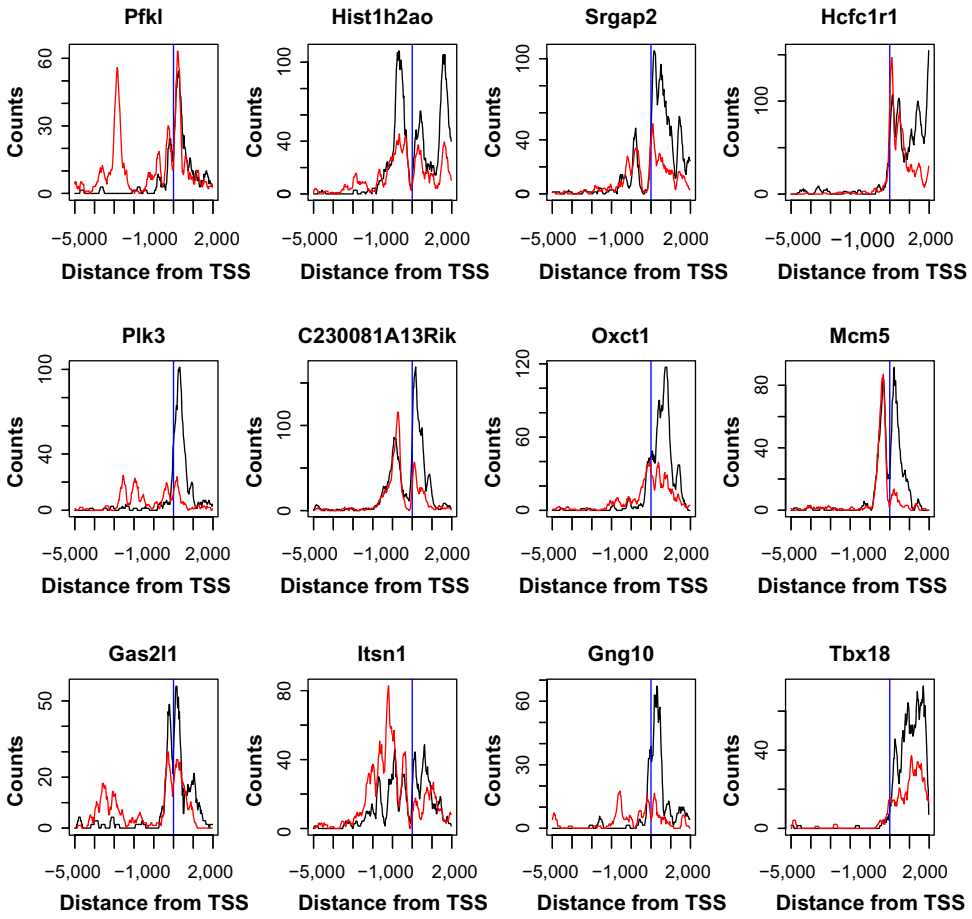

Figure 3. Observed ChIP-seq bin-counts over the promoter region for day –2 (red) and day 7 (black) for 12 genes with large $Z_{0\lambda,WH}$ but small fold changes. Vertical line represents the transcription starting site.

Differential enrichment statistics and gene expression changes. We next investigate the relationship between our test statistics $Z_{0\lambda,WH}$ and changes in expressions of the genes between the two time points. The gene expression data contain two replicates for each time point, and we take the average of two replicates as the mean value W_{ij} for each gene $i = 1, \ldots, m$ and time point $j = 1, 2$. We define the $\log_2$ of the fold change of the expression levels as

$$\Delta W_i = \log_2 \frac{W_{i2}}{W_{i1}}$$

for the ith gene. We then divide the genes into two groups depending on whether higher enrichment was observed at day 7 or day –2. Specifically, we fit the kernel smoothing curve to data for each gene under day 7 and day –2 and obtain the maximum of the curves. The genes are classified as being enriched at day 7 (or day –2) if the maximum height is higher at day 7 (or day –2). Figure 4 shows the gene expression fold changes against the test statistics $Z_{o\lambda,WH}$ together with the Lowess fit for genes that are enriched at day –2. We observe that larger enrichment statistics correspond to down-regulation of these genes. Similarly, Figure 4 also shows the gene expression fold changes against the test statistics $Z_{0\lambda,WH}$ together with Lowess fit for genes that are enriched at day 7. We observe that larger statistics correspond to up-regulation of these genes. Both plots make biological sense since enrichment of H3K27ac is known to promote gene expression. As a comparison, similar plots are shown in Figure 4 for the fold-change statistics. The patterns from the fold-change statistics are not as clear as using our proposed statistics $Z_{0\lambda,WH}$.

To demonstrate this further, we define gene i as being up-regulated if $\Delta W_i > 1$ and down-regulated if $\Delta W_i < -1$. In Figure 5A, we divide our test statistics $Z_{0\lambda,WH}$ into equal-length intervals (<0, 0–5, 5–10, 10–15, 15–20, >20) for the genes that have higher enrichment at day –2. We observe that the proportion of down-regulated genes increases as the test statistics increase. On the other hand, the proportions remain almost constant and close to zero for up-regulated genes. In contrast, for the genes that have higher enrichment at day 7, we observe exactly the opposite (see Fig. 5B). This indicates that our statistics correspond to gene expression changes very well. As a comparison, we present similar plots of the genes based on fold changes of the total reads counts (see Fig. 5C and D). We observe that the separations are not as clear as using our proposed statistics.

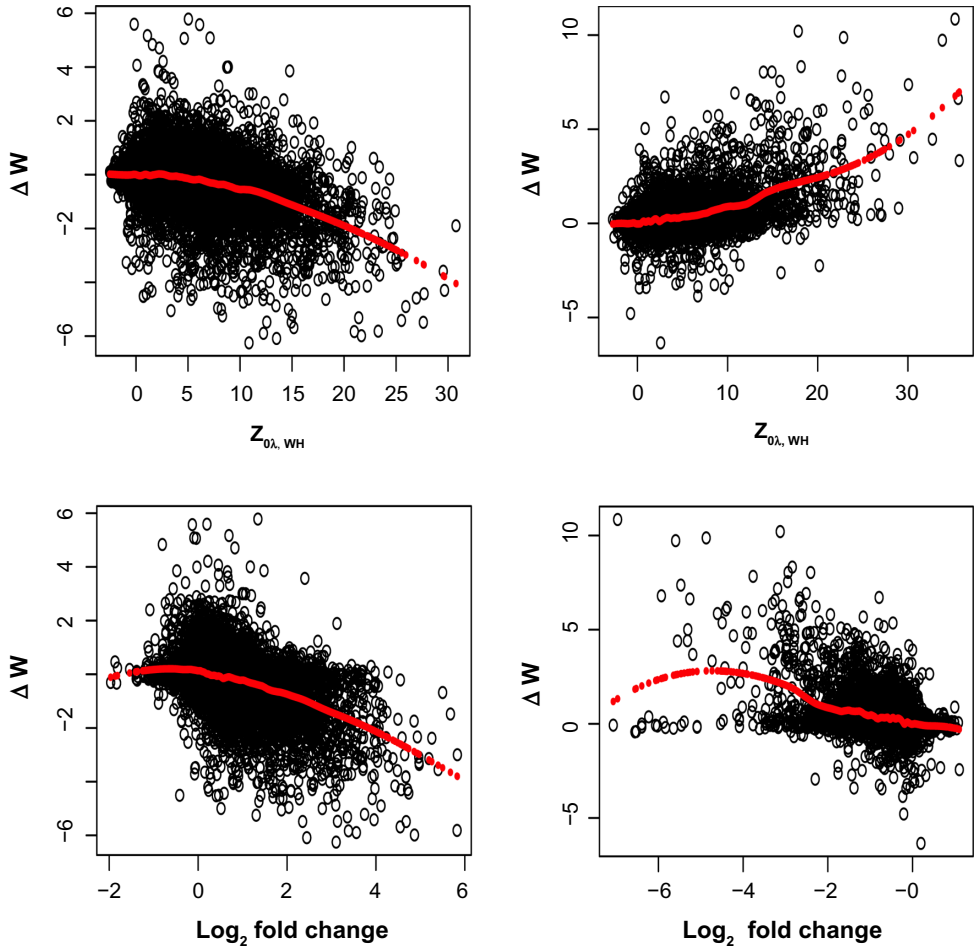

Figure 4. Plots of gene expression fold changes as a function of two different test statistics. Top: proposed smoothing-kernel test statistics; bottom: fold changes. Left panel: genes with enriched H3K27ac binding at day −2; right panel: genes with enriched H3K27ac binding at day 7.

Prediction of gene expression fold changes using histone modification profiles. We next evaluate how well our proposed statistics can be used for predicting the fold changes of gene expression using ChIP-seq data. Besides the H3K27ac ChIP-seq data, we also have data from another five HMs, including H3K4mel, H3K4me2, H3K4me3, H3K27me3, and H3K36me3. In addition, for each gene, besides the promoter region, we also consider the histone modifications in gene body and downstream regions. We evaluate the prediction for fold changes of gene expression by randomly selecting half of the genes as the training set and fit a linear regression model,

$$\Delta W_i = \beta_0 + \sum_{h=1}^{6}\sum_{l=1}^{3}\beta_{hl}TS_{i,hl}, \qquad (11)$$

where h indexes the six HMs and l indexes promoter region, gene body, and down stream region. Using the fitted model, we then predict the gene expression for the left-out genes. We repeat this process 100 times and calculate the average R^2 for model fits for the training genes and the prediction error for genes in the testing sets. As a comparison, we also consider the same model as (11) using the simple fold change statistics as the predictors. Figure 6 shows the model fit for training genes and prediction results for testing genes using our proposed statistics $Z_{0\lambda,WH}$and the fold change statistics as predictors. Clearly, we observe that our proposed statistics give a much better model fit and better prediction results. The average R^2 over 100 random splitting of the genes is 0.57 using our statistics and 0.46 using simple fold changes, and the average prediction error is 0.47 using our statistics and 0.59 using simple fold changes.

We also observe that histone modification dynamics in the promoter and gene body are more predictive than the signals in the downstream regions for predicting the gene expression changes (see Table 1 for details). This is expected since the HMs we used are associated with transcriptional initiation (H3K4me3), open chromatin and cisregulatory activities (H3K4me2/mel and H3K27ac), transcription elongation (H3K36me3), and polycomb-mediated repression (H3K27me3).

The results are based on 100 runs of randomly selected half of the genes as training set and another half as testing set. Numbers in parentheses are standard errors.

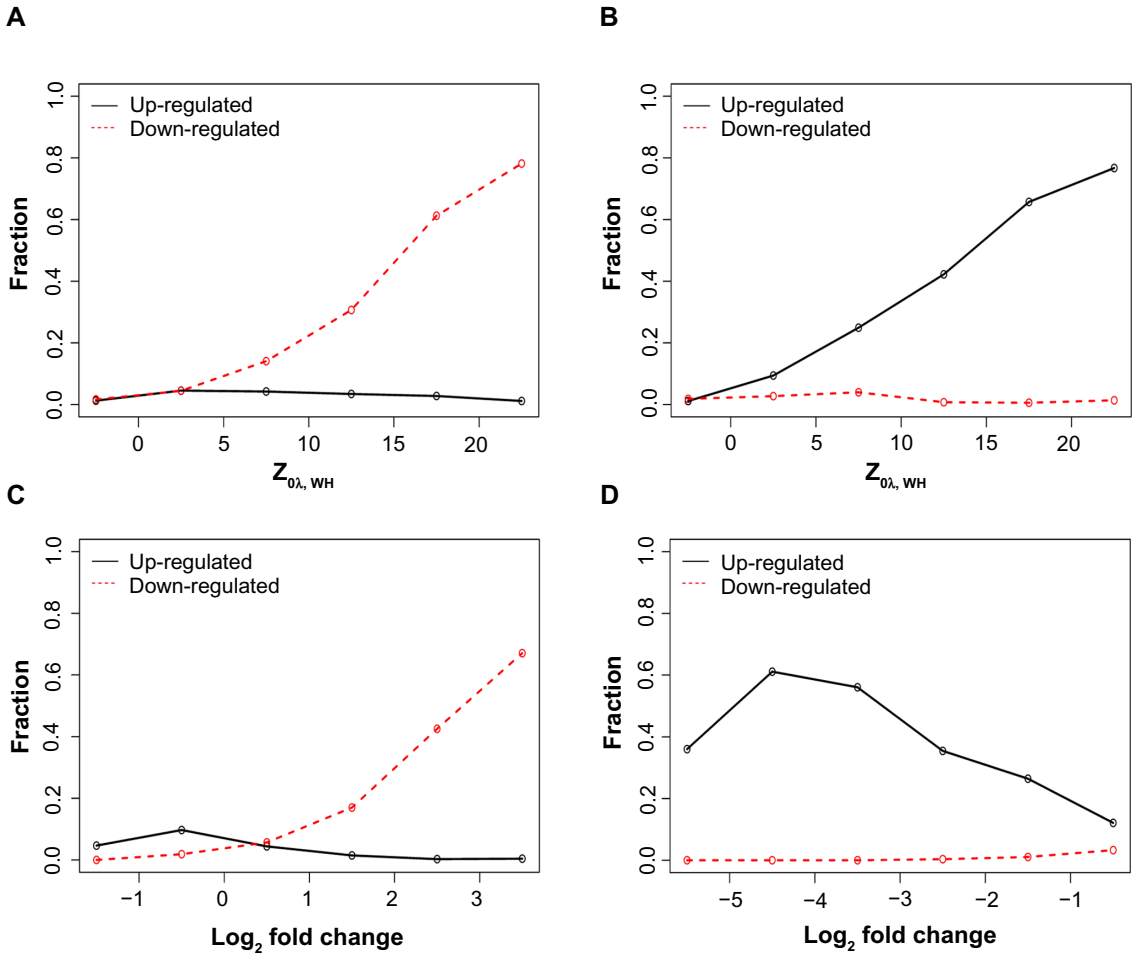

Figure 5. Plots of proportions of up/down-regulated genes in different intervals of the test statistics for the mouse adipogenesis ChIP-seq data, (**A**)–(**B**): proposed smoothing-kernel test statistics; (**C**)–(**D**): fold change statistics; (**A**), (**C**): genes with enriched H3 K27ac at day −2; (**B**), (**D**): genes with enriched H3 K27ac at day 7.

Effects of Bandwidth Selection on Identifying the Differential Enrichment Genes

In applying our kernel-based test in analyzing the mouse ChIP-seq data, we used a global bandwidth of λ = 20/280 for all the genes. Any reasonable test should capture the spatial profiles of signals in the gene regions of interest. On the other hand, the test should also smooth out the small local noises, which are not biologically interesting. We suggest using a relatively large bandwidth to reduce possible false positives. Alternatively, the standard method is to apply cross-validation to find the optimal rate $c(1/n)^{1/5}$.[29] Neumeyer and Dette[30] suggests to obtain the nonparametric variance estimator $\hat{\sigma}_i^2$[31] for each gene and to estimate the bandwidth as follows

$$\lambda = \left\{ \frac{median(\hat{\sigma}_i^2, i = 1, \ldots,)}{n} \right\}^{1/5}.$$

We study the sensitivity of bandwidth selection on the performance of our proposed kernel-based test by considering different bandwidth values, λ_1 = 5/280, λ_2 = 20/280, λ_3 = 60/280, and λ_4 = 90/280. Here, λ_3 and λ_4 correspond to the bandwidths chosen by the nonparametric variance estimation method[30] and the optimal rate $(1/n)^{1/5}$,[29] respectively. We calculate the kernel-based test statistics and denote these statistics as $Z_{\lambda_l, WH}, l = 1, 2, 3, 4$. We present in Figure 7 the histogram of $Z_{\lambda_l, WH}, l = 1, 2, 3, 4$. For the 9,874 genes with the maximum number of read count in both days fewer than 5, which are analogs to the plot in Figure 1B. Clearly, the statistics $Z_{\lambda_1, WH}$ with a relatively small bandwidth lead to false positive detection where the distribution of null genes clearly deviates to the right side of $N(0, 1)$. On the other hand, when a large bandwidth is used, as in statistics $Z_{\lambda_3, WH}$ and $Z_{\lambda_4, WH}$ the tests are conservative, although they still fit the standard normal density curves (red line) reasonably well.

We also examine how different bandwidths affect the ability of identifying differentially expressed genes, where a gene is defined as a true differentially expressed gene if $|\Delta W_i| > 1$. Overall, we observe that it is essential to smooth out the small local signals in order to reduce false-positive identification of genes with differential enrichment. A larger bandwidth gives better results than the smaller ones.

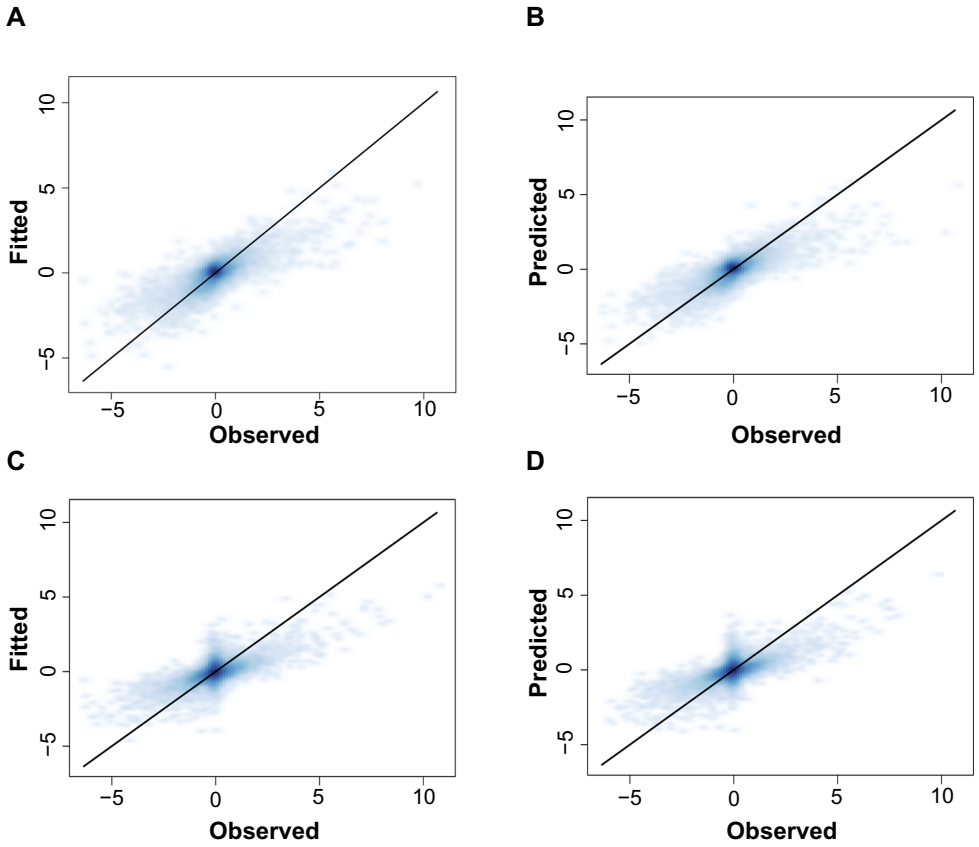

Figure 6. Model fit (left panel) and prediction (right panel) for log of the gene expression fold changes using the proposed statistics $Z_{0\lambda,HW}$ (top panel) and fold changes (bottom panel) of six histone-modification ChIP-seq data in promoter, gene body, and downstream region.

Application to an ENCODE ChIP-Seq Data with Two Replicates

To further evaluate the possible false positives in identifying genes with differential histone modification, we analyze ChIP-seq data reported in the Encyclopedia of DNA Elements (ENCODE) project[32] for a B-lymphoblastoid cell line of human GM12878, which is also part of the 1000 Genomes project, and HeLa-S3 cervical carcinoma cells. Our analysis still focuses on the H3K27ac mark at the promoter regions of the genes with count data available in $n = 280$ bins for each gene. In this experiment, there are a total of $m^* = 23807$ genes. Besides the ChIP-seq data for two biological replicates, two input data are also available. Ideally, we should not expect any genes with differential enrichment between the two replicates. We apply the same procedure as in our analysis of the mouse data in the Application to a Comparative ChIP-seq Study During Mouse Adipogenesis section to the data between two ChIP-seq replicates and calculate test statistics $Z_{new,i}$ for each gene i, $i = 1, \ldots, m = 23{,}807$. The histogram of Z_{new} for all the genes in Figure 8 (top plot) shows that the majority of the test statistics follow the standard normal distribution. In addition, using a Bonferroni-adjusted P value of 0.05, our procedure identifies only 263 genes that show differential enrichment between the two replicates, which results in a less than 1.5% false discovery rate. This analysis further demonstrates that our proposed kernel-based nonparametric testing procedure is not only powerful enough to detect the true differential enriched regions but also makes fewer false identifications.

Table 1. Comparison of model fit R^2 and prediction R^2 (PE) of gene expression fold changes using the proposed statistic $Z_{0\lambda,WH}$ and fold change based on ChIP-seq data of promoter, gene body, and downstream regions of all six HMs as predictors and models using all the three regions.

	$Z_{0\lambda,WH}$		FOLD CHANGE	
	R^2	PE	R^2	PE
Promoter	0.45 (0.009)	0.60 (0.012)	0.35 (0.009)	0.72 (0.015)
Gene body	0.49 (0.008)	0.57 (0.015)	0.40 (0.011)	0.66 (0.014)
Downstream	0.3c0 (0.009)	0.78 (0.018)	0.18 (0.007)	0.90 (0.023)
All regions	0.57 (0.008)	0.47 (0.013)	0.46 (0.009)	0.59 (0.012)

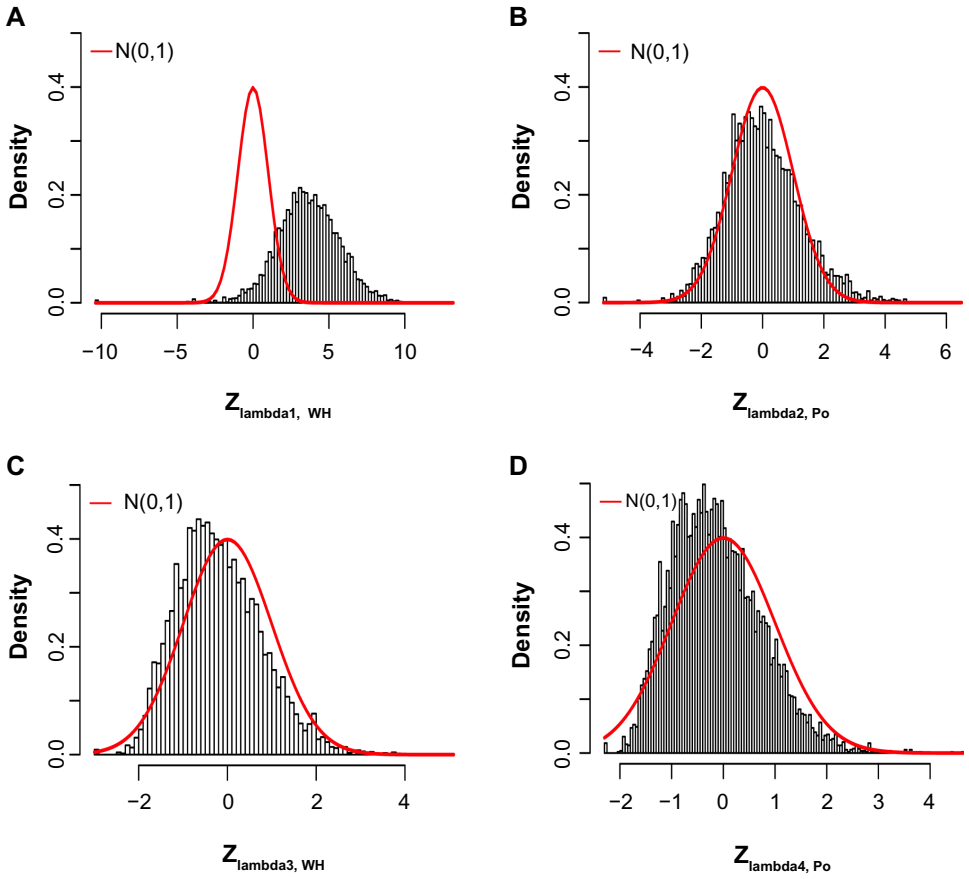

Figure 7. Histogram of the test statistics $Z_{\lambda_i,WH}$ with the different bandwidths: (**A**) $\lambda_1 = 5/280$, (**B**) $\lambda_2 = 20/280$, (**C**) $\lambda_3 = 60/280$, (**D**) $\lambda_4 = 90/280$ for 9,874 genes with the maximum number of read count in both day −2 and day 7 fewer than 5 in mouse adipogenesis ChIP-seq data.

We also perform an analysis to identify the genes with differential enrichment of histone modification between a B-lymphoblastoid cells and HeLa-S3 cervical carcinoma cells. Figure 8 (bottom plot) shows the histogram of the test statistics for all 23,807 genes. Using a Bonferroni threshold for genome-wide level of 0.05, we identify 6,647 genes that show differential H3K27ac enrichment at their promoter regions.

Conclusions and Discussion

We have proposed a kernel-smoothing-based nonparametric test to identify genes with differential histone enrichment for ChIP-seq data. Different from all the currently available methods, our method models the spatial histone enrichment profiles at the promoter regions of the genes, rather than simply modeling the total read counts in a given window. The method can therefore capture different types of differences in protein-enriched profiles between two experimental conditions. To detect differences in enrichment profiles, we constructed a nonparametric statistic based on kernel smoothing on the differences of the profiles after approximate normal transformation of the data. We have shown that the proposed test statistic corresponds to the gene expression changes better than other statistics and the models based on a combination of different HMs can effectively predict the gene expression fold changes. Although prediction of gene expression using the ChIP-seq data has been studied in many published works,[33,34] these papers focused only on prediction of gene expression at a static state. Our results further demonstrate that change of histone modifications and the dynamic chromatin signatures can also be very predictive for the fold-changes of gene expression between two different cellular states.

We considered only the problem of identifying the differential enrichment regions between two conditions, where we fit the kernel-smoothing to the differences of the normal transformed data in order to further smooth out the small local changes that might be due to differences in GC contents or mappability of the sequencing reads. By smoothing, we expect that our procedure is robust to such small changes due to genomic features. If input data are available, one can take the difference of the square-root transformed count data between ChIP and input and then apply our proposed test with kernel smoothing. The method requires the users to specify the regions to test. Besides the regions close to genes as we tested in this paper, one can also first identify the histone-enriched regions using some existing methods such as MACS[7] or SICER[35] and then test for differential enrichments using our proposed methods. Finally, our proposed method can be extended to identify differential enrichment in multiple conditions. In such cases, we can define the test statistic as the mean or maximum of all the pair-wise statistics as proposed in this paper.

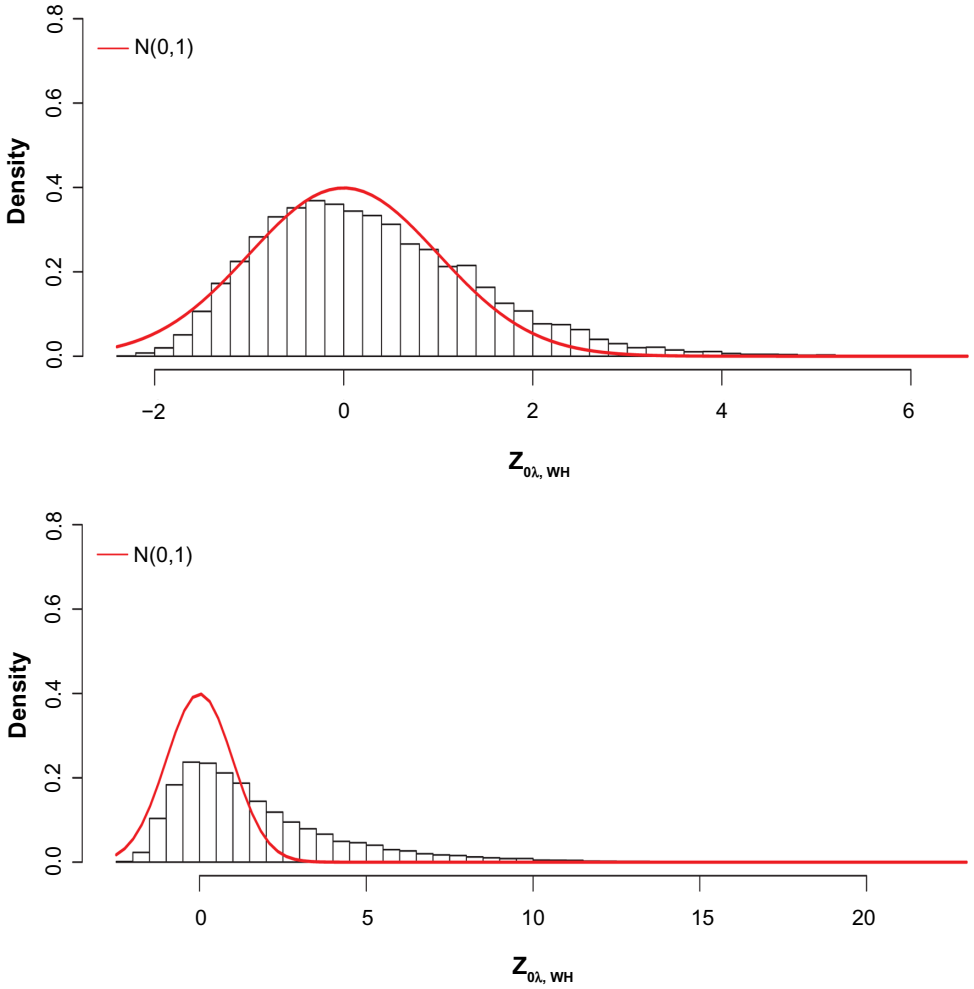

Figure 8. Top: Histogram of differential enrichment test statistics Z_{new} between two biological replicates of the ENCODE data for all 23,807 genes. Bottom: Histogram of differential enrichment test statistics Z_{new} between two cell types (B-lymphoblastoid cell vs HeLa-S3 cervical carcinoma cells) of the ENCODE data for all 23,807 genes. The red curve represents the standard normal density.

Author Contributions

Conceived and designed the experiments: HL, QW, KW. Analyzed the data: HL, QW, KW. Wrote the first draft of the manuscript: QW, HL. Contributed to the writing of the manuscript: HL, QW, KW. Agree with manuscript results and conclusions: QW, KW, HL. Jointly developed the structure and arguments for the paper: QW, HL, KW. Made critical revisions and approved final version: QW, KW, HL. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Park P. ChIP-Seq: advantages and challenges of a maturing technology. *Nat Rev Genet*. 2009;10:669–80.
2. Johnson D, Mortazavi A, Myers R, Wold B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*. 2007;316:1497.
3. Mikkelsen T, Xu Z, Zhang X, et al. Comparative epigenomic analysis of murine and human adipogenesis. *Cell*. 2010;143:156–69.
4. Mortazavi A, Williams B, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 2008;5:621–8.
5. Barski A, Cuddapah S, Cui K, et al. High-resolution profiling of histone methylations in the human genome. *Cell*. 2007;129:823–37.
6. Hon G, Wang W, Ren B. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput Biol*. 2009;5:el000566.
7. Zhang Y, Liu T, Meyer C, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9:R137.
8. Kuan P, Chung D, Pan G, Thomson J, Stewart R, Kele S. A statistical framework for the analysis of ChIP-Seq data. *J Am Stat Assoc*. 2011;106:891–903.
9. Ji H, Jiang H, Ma W, Johnson D, Myers R, Wong W. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol*. 2008;26:1293–300.
10. Schwartzman A, Jaffey A, Gavrilovz Y, Meyer C. Multiple testing of local maxima for detection of peaks in ChIP-Seq data. *Ann Appl Stat*. 2013;7:471–94.
11. Spyrou C, Stark R, Lynch A, Tavaré S. BayesPeak: Bayesian analysis of ChIP-seq data. *BMC Bioinformatics*. 2009;10:299.
12. O'Geen H, Echipare L, Farnham P. Using ChIP-seq technologe high-resolution profiles of histone modifications. *Methods Mol Biol*. 2011;791:265–86.
13. Liang K, Keles S. Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics*. 2012;28:121–2.
14. Chen Y, Jrgensen M, Kolde R, et al. Prediction of RNA polymerase II recruitment, elongation and stalling from histone modification data. *BMC Genomics*. 2011;12:544.
15. He H, Meyer C, Shin H, et al. Nucleosome dynamics defines transcriptional enhancers. *Nat Genet*. 2010;42:343–7.
16. Angel A, Song J, Dean C, Howard M. A polycomb-based switch underlying quantitative epigenetic memory. *Nature*. 2011;476:105–8.
17. Stark R, Brown G. DiffBind: differential binding analysis of ChIP-Seq peak data. *Bioconductor*. 2011. http://bioconductor.org/packages/release/bioc/vianettes/DiffBind/dec/DiffBinded/.
18. Taslim C, Wu J, Yan P, et al. Comparative study on ChIP-seq data: normalization and binding pattern characterization. *Bioinformatics*. 2009;25:2334–40.

19. Shao Z, Zhang Y, Yuan G, Orkin S, Waxman D. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biol*. 2012;13:R16.
20. Langmead B, Trapnell C, Pop M, Salzberg S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10:R25.
21. Brown L, Cai T, Zhang R, Zhao L, Zhou H. The root-unroot algorithm for density estimation as implemented via wavelet block thresholding. *Probab Theory Relat Fields*. 2010;146:401–33.
22. Brown L, Gans N, Mandelbaum A, et al. Statistical analysis of a telephone call center: a queing science perspective. *J Am Stat Assoc*. 2005;100:36–50.
23. Lepski O, Spokoiny V. Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative. *Bernoulli*. 1999;5:333–58.
24. Hastie T, Tibshirani R. *Generalized Additive Models*. Vol 43. Chapman and Hall/CRC; 1990. London.
25. Bentler P, Xie J. Corrections to test statistics in principal Hessian directions. *Stat Probab Lett*. 2000;47:381–9.
26. Wilson E, Hilferty M. The distribution of chi-squared. *Proc Natl Acad Sci U S A*. 1931;17:684–8.
27. Efron B, Tibshirani R, Storey J, Tusher V. Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc*. 2001;96:1151–60.
28. Tusher V, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*. 2001;98:5116–21.
29. Gasser T, Kneip A, Köhler W. A flexible and fast method for automatic smoothing. *J Am Stat Assoc*. 1991;86:643–52.
30. Neumeyer N, Dette H. Nonparametric comparison of regression curves: an empirical process approach. *Ann Stat*. 2003;31:880–920.
31. Rice J. Bandwidth choice for nonparametric regression. *Ann Stat*. 1984;12:1215–30.
32. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489:57–74.
33. Karlic R, Chung H, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A*. 2010;107:2926–31.
34. Dong X, Greven M, Kundaje A, et al. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol*. 2012;13:R53.
35. Zang C, Schones D, Zeng C, Cui K, Zhao K, Peng W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*. 2009;25:1952–8.